Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Borut Flis

Napovedovanje dobička pri športnih stavah

diplomsko delo

visokošolski strokovni študijski program prve stopnje računalništvo in informatika

Mentor: viš. pred. dr. Aleksander Sadikov

Ljubljana 2014

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

[1] V dogovorju z mentorjem lahko kandidat diplomsko delo s pripadajočo izvorno kodo izda tudi pod katero izmed alternativnih licenc, ki ponuja določen del pravic vsem: npr. Creative Commons, GNU GPL.

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Diplomska naloga se ukvarja z napovedovanjem rezultatov košarkarskih tekem. Cilj je ustvariti napovedni model, ki bi bil dovolj natančen, da bi lahko pozitivno stavil na športnih stavnicah. V diplomski nalogi je preizkušenih več algoritmov strojnega učenja, ki se učijo na podatkih tekem redne sezone ameriške profesionalne košarkaške lige (NBA). Napovedni model vrne verjetnost zmage izbrane ekipe v določeni tekmi. Vrnjene verjetnosti se uporabijo za analizo tveganja. Najboljši model se nato uporabi na stavnici in izmeri potencialni dobiček.

Izjava o avtorstvu diplomskega dela

Spodaj podpisani Borut Flis, z vpisno številko **63100409**, sem avtor diplomskega dela z naslovom:

Napovedovanje dobička pri športnih stavah

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom viš. pred. dr. Aleksandra Sadikovega,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 18. septembra 2014 Podpis avtorja:

Zahvaljujem se mentorju Aleksandru Sadikovemu in svoji družini za podporo.

Kazalo

Povzetek

V diplomski nalogi se ukvarjamo z napovedovanjem rezultatov košarkarskih tekem. Želimo si ustvariti napovedni model, ki bo dovolj natančen, da bomo z njim lahko zaslužili denar na športnih stavnicah. Ena tekma v redni sezoni lige NBA je en učni primer. Vsak primer ima več atributov, ki opisujejo nasprotujoči ekipi. Preizkusili smo veliko statističnih atributov, za katere smo verjeli, da so pomembni za napoved zmagovalca. Preizkusili smo različne napovedne modele, ki bodo vračali verjetnost domače in verjetnost gostujoče zmage. Vrnjene verjetnosti bomo uporabili za analizo tveganja. Najboljši model smo uporabili na stavnicah in izmerili potencialni dobiček. Ugotovitve te diplomske naloge bi bile lahko koristne tudi na drugih področjih, ki se ukvarjajo s tveganjem in napovedovanjem prihodnosti. Rezultati te naloge niso popolni, v zaključku smo omenili še možne izboljšave.

Ključne besede

strojno učenje, športne stave, košarka, NBA, analiza tveganja

Abstract

We wish to build a model, which could predict the outcome of basketball games. The goal was to achieve an sufficient enough accuracy to make a profit in sports betting. One learning example is a game in the NBA regular season. Every example has multiple features, which describe the opposing teams. We tried many methods, which return the probability of the home team winning and the probability of the away team winning. These probabilities are used for risk analysis. We used the best model in hypothetical sports betting and measured potential net profit. The results are not perfect, so we mentioned possible improvements. I think that a lot of the ideas could also be of use in other fields, which deal with risk and predicting the future.

Keywords

machine learning, sport betting, basketball, NBA, risk analysis

Poglavje 1

Uvod

1.1 Motivacija

V zadnjem času se zelo povečuje uporaba statistike in matematike v športu. Najbolj je to področje razvito v ameriških profesionalnih ligah. Športno statistiko uporabljajo trenerji, ekonomisti in aktuarji. Podatki so idealni za uporabo strojnega učenja in podatkovnega rudarjenja. Športne stave so velik posel. Stavnice potrebujejo analitike, ki izračunajo verjetnosti dogodkov in izberejo kvote, ki stavnici dolgoročno omogočajo dobiček. V tej diplomski nalogi nas športne stave zanimajo iz perspektive uporabnika stavnic. Zanima nas, kako izbrati prave stave, ki bi dolgoročno prinašale dobiček. Stavnice kvote prilagodijo navadam uporabnikov; ko uporabniki vložijo veliko denarja na določen izid, se kvota zmanjša, da se zmanjša tveganje. Poizkušali bomo narediti model, ki bo pametnejši od kvot. Pridobljena znanja bodo uporabna tudi iz perspektive stavnice.

1.2 Kratek opis problema

V diplomski nalogi se bomo ukvarjali z napovedovanjem zmagovalca košarkarskih tekem. S pomočjo teh napovedi bomo poskušali ustvariti dobiček na športnih stavnicah. Cilj naloge je torej narediti program, ki bo z zadostno točnostjo

napovedoval rezultate košarkarskih tekem, da bi lahko ustvarili dobiček na stavnicah. En učni primer je ena tekma v redni sezoni košarkarske lige NBA. Da se modeli v učenju ne bodo preveč prilagodili specifičnim primerom, je treba zbrati skupaj čim večje število primerov. Na podatkih bomo preizkusili več metod. Napovedni model bo vrnil verjetnosti zmage domače in zmage gostujoče ekipe, te verjetnosti bomo zmnožili s kvotami, ki jih ponujajo stavnice, ter izračunali pričakovani dobiček. Na voljo imamo tudi za dve sezoni podatkov kvot iz športnih stavnic, na njih bomo preizkusili najboljšo metodo in ocenili, kakšen bi bil hipotetični dobiček, če bi na tekme s pozitivno pričakovano vrednostjo stavili.

1.3 Uporabljena orodja

Vso kodo smo napisali v programskem jeziku Python. Uporabljali smo tudi knjižnico Orange. Nekatere algoritme smo sami spisali v Pythonu, pri nekaterih pa smo uporabili Orangeovo implementacijo.

Poglavje 2

Opis podatkov

2.1 Opis učnega primera

En učni primer je ena tekma v redni sezoni ameriške košarkarske lige NBA. Razredni atribut je zmagovalec tekme, ostali atributi pa so povprečje statističnih podatkov prejšnjih tekem. Podatke, ki smo jih uporabili v diplomski nalogi, smo pridobili na strani http://www.basketball-reference.com/, kjer so javno dostopni v HTML-ju. Na voljo je več sezon, da bi preprečili pretirano prilagajanje podatkom, smo zbrali primere iz kar osmih sezon. Zapisnik tekme je podan v obliki tabele(ang. box score), primer je viden na sliki 2.1. Podatke smo pridobili z iteracijo po tekmah v sezoni in potem razčlenjevanju HTML-ja zapisnikov. V podatkovnih strukturah hranimo vrednosti atributov prejšnjih tekem. Podatke iz zapisnika dodamo prejšnjim vrednostim šele potem, ko že zapišemo učni primer, saj morajo biti atributi, ki jih uporabljamo za napovedovanje dostopni že pred tekmo.

2.2 Opis atributov

Razmerje zmag v medsebojnih dvobojih
V podatkih je razmerje zmag v medsebojnih dvobojih izraženo kot odstotek zmag ekipe, ki igra doma. Upoštevanih je zadnjih deset medsebojnih

Washington Wizards (2-4) Glossary · SHARE · Embed · CSV · Export · PRE · LINK · ?

	Basic Box Score Stats																		
Starters	**MP**	**FG**	**FGA**	**FG%**	**3P**	**3PA**	**3P%**	**FT**	**FTA**	**FT%**	**ORB**	**DRB**	**TRB**	**AST**	**STL**	**BLK**	**TOV**	**PF**	**PTS**	**+/-**
Bradley Beal	42:11	13	23	.565	6	8	.750	2	2	1.000	1	5	6	0	1	0	1	0	34	0
Trevor Ariza	40:49	7	13	.538	1	2	.500	0	5	.000	1	4	5	2	0	0	1	3	15	+3
John Wall	39:40	3	13	.231	1	5	.200	3	3	1.000	0	5	5	8	0	1	4	0	10	-1
Marcin Gortat	39:00	4	9	.444	0	0		3	5	.600	0	8	8	2	1	4	1	2	11	-4
Nene Hilario	25:21	5	6	.833	0	0		4	10	.400	3	4	7	6	0	1	0	3	14	+4
Reserves	**MP**	**FG**	**FGA**	**FG%**	**3P**	**3PA**	**3P%**	**FT**	**FTA**	**FT%**	**ORB**	**DRB**	**TRB**	**AST**	**STL**	**BLK**	**TOV**	**PF**	**PTS**	**+/-**
Martell Webster	29:10	3	10	.300	1	7	.143	1	1	1.000	1	4	5	0	1	1	1	4	8	-7
Al Harrington	22:21	4	10	.400	3	6	.500	0	0		0	1	1	2	0	0	2	2	11	-7
Eric Maynor	13:20	0	1	.000	0	1	.000	0	0		0	2	2	2	1	0	2	2	0	0
Kevin Seraphin	13:08	1	3	.333	0	0		0	0		1	1	2	1	1	0	0	3	2	+7
Trevor Booker	*Did Not Play*																			
Jan Vesely	*Did Not Play*																			
Glen Rice	*Did Not Play*																			
Garrett Temple	*Did Not Play*																			
Team Totals	**265**	**40**	**88**	**.455**	**12**	**29**	**.414**	**13**	**26**	**.500**	**7**	**34**	**41**	**23**	**5**	**7**	**12**	**19**	**105**	

Slika 2.1: Zapisnik tekme Washington - Oklahoma City 10. 11. 2013 na basketball-reference.com

dvobojev, kar presega eno sezono, ekipe se v časovnem obdobju teh desetih dvobojev tudi spremenijo, vendar obstajajo ekipe, ki so lahko določeni ekipi neugodne v daljšem časovnem obdobju.

Rezultati v zadnjem času

Beležimo odstotek zmag domače in gostujoče ekipe na zadnjih 10 tekmah.

Prosti dnevi

Za nasprotujoči ekipi beležimo število prostih dni pred tekmo. Liga NBA ima zelo natrpan urnik, bolj kot ostala športna prvenstva, v redni sezoni vsako moštvo v manj kot pol leta odigra 82 tekem. Ekipe včasih igrajo več dni zapored, včasih pa imajo na voljo nekaj prostih dni, ki jih lahko izkoristijo za boljšo pripravo na nasprotnika in počitek. Izkaže se, da je to zelo pomemben atribut.

Število zaporednih tekem v gosteh

V ligi NBA se večkrat zgodi, da določena ekipa igra več tekem zapored v gosteh, kar pomeni veliko utrujajočih potovanj in malo možnosti za trening

ter pripravo na nasprotnika. Zato sklepamo, da je za gostujočo ekipo bolje, če je odigrala manj tekem zapored v gosteh.

Pace

Angleško poimenovanje statističnega podatka, ki ocenjuje število posesti določene ekipe na tekmi.

$$\frac{48 * (\text{število posesti ekipe} + \text{število posesti nasprotnika})}{2 * 48}$$

Skoki

Povprečno število skokov na tekmo.

Dosežene točke

Povprečno število doseženih točk na tekmo.

Prejete točke

Povprečno število prejetih točk na tekmo.

2.3 Neuspeli poskusi novih atributov

Manjkajoči igralci

Poškodba enega ali celo več igralcev lahko zelo vpliva na predstavo ekipe. Atribut manjkajočih igralcev bi moral predstavljati tudi, kako pomembni so ti igralci za ekipo. Če je poškodovan igralec z majhno minutažo, to nima takega vpliva, kot če je poškodovan najboljši igralec. Naša ideja je bila, da bi atribut bil razmerje med povprečnim številom točk poškodovanih igralcev in povprečnim številom točk ekipe. Za vsako tekmo v preteklosti vemo, kateri igralci so igrali. Za naslednjo tekmo pa ne moremo biti prepričani. V medijih so večkrat nasprotujoče si informacije o poškodbah določenih igralcev, veliko je izmišljenih govoric, včasih ekipe tudi namerno zavajajo o poškodbah, saj želijo zmesti nasprotnika. Znano je tudi, da nekatere ekipe najboljšim igralcem včasih namenijo počitek [2]. V realni situaciji pred tekmo je najbrž mogoče z določeno verjetnostjo oceniti, kateri igralci bodo igrali, saj vemo, kateri mediji so bolj zanesljivi, in vemo, da so morda kakšni igralci že dalj časa poškodovani. Pri učnih primerih izpred nekaj sezon pa je to zelo težko

narediti. Če bi v učnih primerih dodali atribut manjkajočih igralcev na podlagi zapisnikov tekem, bi s tem pridobili nepravično prednost.

Koliko zadnjih tekem upoštevamo

Eden izmed atributov so rezultati v zadnjem času. Vprašanje, ki se zastavlja, je, koliko zadnjih tekem je najbolje upoštevati. Če vzamemo več tekem, s tem zmanjšamo faktor sreče, vendar ne upoštevamo dejstva, da so se lahko v času zadnjih nekaj tekem zgodile poškodbe ali podobne nevšečnosti, ki vplivajo na predstave ekip. Razmišljali smo o tem, da dobre ekipe včasih izgubijo kakšno tekmo, slabe ekipe pa včasih kakšno zmagajo, vendar redko se zgodi, da tak niz traja dalj časa. Problem je izključujoči ali (XOR); če je ekipa dobra, vendar je zadnjo tekmo izgubila, ima veliko možnosti za zmago, podobno kot so možnosti za poraz velike, če je ekipa slaba, vendar je zadnjo tekmo izgubila. Po testiranju se je izkazalo, da ta ideja ne prinaša dobrih rezultatov. Točnost napovedi je približno enaka ne glede na to, koliko zadnjih tekem upoštevamo.

Pomembnost tekem

Vsaka ekipa v redni sezoni odigra 82 tekem. Cilj vsake ekipe je uvrstitev v končnico, poleg tega pa višja uvrstitev prinaša tudi boljše izhodišče. Problem pa je v tem, da vse ekipe niso vedno motivirane, da dosežejo čim več zmag. Slabša uvrstitev na lestvici namreč prinaša boljšo pozicijo na naslednjem naboru, kjer ekipe izbirajo igralce. Govorice so, da nekatere ekipe zaradi tega namerno izgubljajo[1]. Vendar gre bolj za špekulacije kot dejstva, ne moremo biti prepričani, da določena ekipa namerno izgublja. Pomembnost tekme in motivacijo je težko izraziti kot statistični atribut.

2.4 Pomembnost atributov

Postavlja se vprašanje, kateri atribut je najpomembnejši. Pri vsakem atributu beležimo vrednost za domačo in gostujočo ekipo. Preverjali smo, kateri atribut bolje napove zmagovalca, tako da smo primerjali atribut domače in gostujoče ekipe. Če določen atribut s približno polovično natančnostjo na-

Atribut	Točnost napovedi
Rezultati v zadnjem času	0,637
Število točk	0,598
Prosti dnevi	0,586
Število prejetih točk	0,560
Skoki	0,516
Pace	0,513

Tabela 2.1: Pomembnost atributov

pove zmagovalca, je najbrž nepomemben. V tabeli 3.3 so vsi atributi in njihova natančnost, če jih uporabljamo samostojno za napoved zmagovalca, razvrščeni so od najboljšega do najslabšega. Presenetljivo dobro se je odrezal atribut prostih dni, to je verjetno posledica izjemno natrpanega urnika v redni sezoni. Zanimivo rezultati v zadnjem času dajo boljši rezultat kot povprečno število zadetih točk ali prejetih točk. To pomeni, da je trenutna forma bolj relevantna od kvalitete čez daljše obdobje. Število prejetih in danih točk bi lahko bil en sam atribut in sicer razlika v točkah, vendar smo se odločili za dva atributa, saj en izraža kvaliteto napada, drugi pa kvaliteto obrambe, v skupnem atributu se ta lastnost izgubi. Ni presenečenje, da je pace precej nepomemben, saj gre za atribut, ki bolj kot kvaliteto izraža slog ekipe.

Poglavje 3

Gradnja in testiranje napovednega modela

3.1 Implementacija napovednih modelov

Odločili smo se, da bomo uporabili nekaj že narejenih implementacij klasifikatorjev, obenem pa bomo zaradi boljšega razumevanja naredili tudi nekaj svojih implementacij. Najprej smo se lotili implementacije naivnega Bayesa. Ta klasifikator smo si izbrali, ker kot rezultat daje verjetnost, kar je idealno za odločanje s tveganjem. Polega tega so pri učnih primerih velikokrat neznane vrednosti in tudi pri uporabi primerov z neznano vrednostjo ima prednost naivni Bayes. Formula za izračun verjetnosti določenega razreda je prikazana (3.1).

$$p(c|v_1, v_2, \ldots v_n) = p(c) * \prod_i \frac{p(c|v_i)}{p(c)} \tag{3.1}$$

Vsi atributi so zvezni, zato moramo uporabiti funkcijo za zvezne atribute (3.2). Ta funkcija predpostavlja, da so vrednosti atributov normalno porazdeljene.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3.2}$$

Pri ostalih metodah smo uporabili Orangeovo implementacijo. Iz knjižnjice classification smo uporabili naslednje algoritme: logistična regresija(logreg),

	Klasifikacijska točnost	Brierjeva mera
Naivni Bayes	0,651	0,451
Logistična regresija	0,661	0,428
Nevronske mreže	0,629	0,470
Odločitveno drevo	0,579	0,555
SVM	0,590	0,476
K-najbližjih sosedov	0,627	0,454

Tabela 3.1: Rezultati testiranja

nevronske mreže(neural), odločitveno drevo(tree), SVM(svm) in K-najbližjih sosedov(knn).

3.2 Način in rezultati testiranja

Testirali smo več različnih metod strojnega učenja. Testirali smo klasifikacijsko točnost in Brierjevo mero. Uporabili smo petkratno prečno preverjanje. Prečno preverjanje smo implementirali z iteracijo čez vse primere, katerim smo dodali ostanek po deljenju njihovega indeksa s številom preverjanj. Ta ostanek bo ob razdelitvi podatkov na učno in testno povedal v katerem preverjanju je določen podatek v testni množici. V našem primeru, kjer verjetnosti, ki jih napove model, uporabljamo pri odločanju s tveganjem, je morda Brierjeva mera še pomembnejša od klasifikacijske točnosti.

$$\text{Brierjeva mera} = \frac{1}{N}\sum_{i=1}^{N}\Big(Napoved_i - Razred_i\Big)^2 \tag{3.3}$$

Če določen klasifikator napove zmagi ekipe verjetnost 60 %, je možnost, da bomo stavili na to ekipo, precej manjša, kot če bi bila napovedana verjetnost 70 %. Brierjeva mera bo kaznovala modele, ki bodo pretiravali z visokimi verjetnostmi.

	Klasifikacijska točnost
Naivni Bayes	0,672
Logistična regresija	0,671
Odločitveno drevo	0,618
SVM	0,604
K-najbližjih sosedov	0,657

Tabela 3.2: Rezultati testiranja s selekcijo atributov

3.3 Izboljšanje modelov s selekcijo atributov

Najboljšo klasifikacijsko točnost doseže logistična regresija. Najboljši prejšnji poskusi napovedovanja zmagovalcev košarkarskih tekem dosežejo točnost okoli 70 % [5]. Naš najboljši model je s 66 % še precej slabši. Upoštevati moramo tudi, da že večinski klasifikator doseže okoli 60 %, saj je to odstotek zmag domačih ekip. Točnost bomo poskušali izboljšati s selekcijo atributov, saj so nekateri preizkušeni modeli zelo slabi pri prepoznavanju nepomembnih atributov, to velja predvsem za naivni Bayes. Uporabljali bomo selekcijo atributov z dodajanjem. Gre za postopek dodajanja atributov z namenom izboljšanja točnosti. Postopek začnemo brez atributov, potem dodamo atribut, ki najbolj poveča točnost, tako nadaljujemo, dokler lahko z dodajanjem še povečujemo točnost[4]. Med postopkom smo točnost preverili s prečnim preverjanjem. Rezultati so se izboljšali. Naš najboljši model je naivni Bayes z atributi: domači rezultati v zadnjem času, razmerje zmag v medsebojnih dvobojih, gostujoči rezultati v zadnjem času, domači skoki, domači prejete točke, domači dosežene točke, gostujoči prejete točke in domači prosti dnevi.

3.4 Sprememba uporabe atributov

V učnih primerih za vsak atribut hranimo vrednost domače in gostujoče ekipe. Poizkušali smo tudi kako se modeli obnesejo, če so atributi razlika med vrednostjo domače in gostujoče ekipe. S tem so atributi izraženi rela-

Verjetnost	Klasifikacijska točnost
0,5	0,490
0,6	0,544
0,7	0,618
0,8	0,709
0,9	0,876

Tabela 3.3: Klasifikacijska točnost pri določeni verjetnosti

Senzitivnost	Specifičnost
0,771	0,515

Tabela 3.4: Senzitivnost in specifičnost

tivno glede na nasprotnika. To je lahko dobro, saj nas pri napovedovanju tekem zanima predvsem kako kvalitetna je določena ekipa v primerjavi z nasprotnikom. Slaba stran pa je, da izgubimo določeno informacijo. Naredili smo nove učne primere in na njih testirali vse modele, izvedli smo tudi selekcijo atributov. Izkaže se, da se rezultat malenkost, najboljši klasifikator je še vedno naivni Bayes, njegova točnost pa doseže 67,77 %. To je najboljši klasifikator, tega bomo uporabili tudi na športnih stavnicah.

3.5 Dodatna analiza modela

Verjetnosti, ki jih vrne model, bomo uporabili kot vhod za odločitveno drevo. Pomembno je, da so te verjetnosti čim bolj podobne točnosti pri taki verjetnosti. Po selekciji atributov je bil kot najboljši model izbran naivni Bayes. Znano je, da ta klasifikator včasih pretirava z visokimi verjetnostmi[6], zato smo izmerili, kakšna je točnost napovedi pri določenih verjetnostih. Kot vidimo v tabeli 3.3, točnost narašča z višjo verjetnostjo. Vseeno pa je točnost manjša od verjetnosti, kar pomeni, da model malenkost pretirava. V tabeli 3.4 sta izmerjeni senzitivnost in specifičnost. Precej večja je specifičnost,

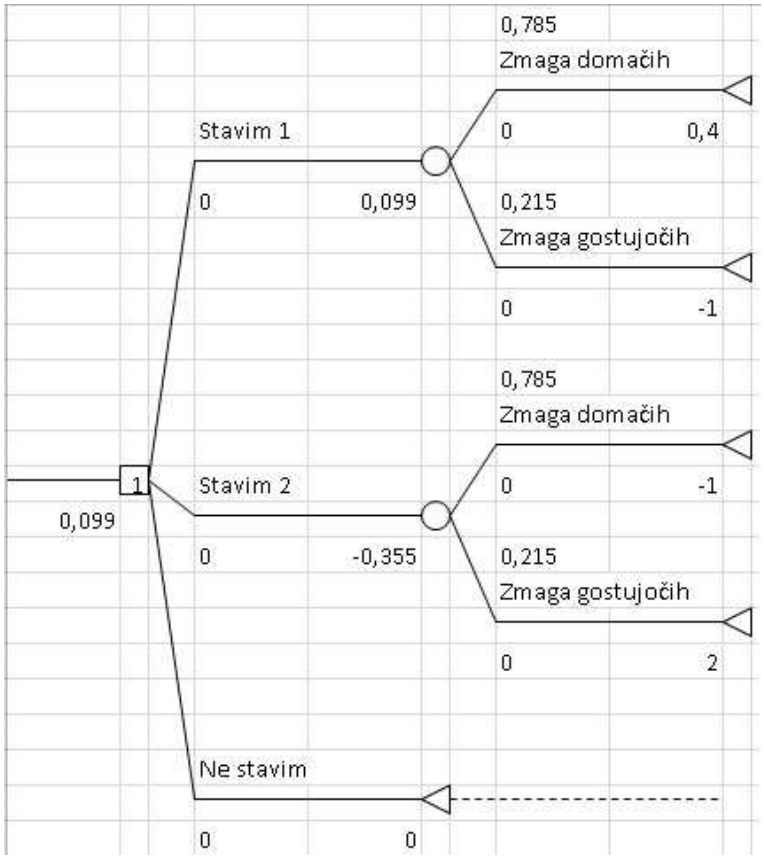
kar pomeni, da je pri domačih zmagah večja točnost. To ni naključje, saj so domače zmage večinski razred.

Poglavje 4

Odločanje s tveganjem in uporaba na stavnicah

4.1 Odločanje s tveganjem

Cilj diplomske naloge je ustvariti model, s katerim bi lahko ustvarili zaslužek na stavnicah. Poleg dobrega napovednega modela potrebujemo tudi natančno oceno pričakovanega dobička in izgube. Napovedni model vrne verjetnosti zmage domače in gostujoče ekipe. Stavnice za vsako tekmo ponujajo kvote za oba možna izida. Pričakovani dobiček izračunamo s pomočjo odločitvenih dreves. Pred vsako tekmo imamo na voljo tri alternative - lahko stavimo na zmago domače ali gostujoče ekipe, lahko pa se tudi odločimo, da ne bomo stavili. Tekma predstavlja dogodek v odločitvenem drevesu, rezultat tekme pa je končno vozlišče. Pričakovana vrednost za končno vozlišče je seštevek zmnožka verjetnosti z vplačilom pomnožene kvote, od katere odštejemo vplačilo, in zmnožka negativnega vplačila ter obratne verjetnosti. Odločimo se za stavo z večjo pričakovano vrednostjo, razen v primeru, kadar sta obe negativni, takrat ne stavimo.



Slika 4.1: Odločitveno drevo za tekmo Memphis - OKC, dne 28. 2. 2014

4.2 Testiranje modelov na stavnicah

Na voljo imamo za dve sezoni kvot. Na teh dveh sezonah bomo testirali, kako bi se naš model obnesel na stavnicah. Na sezonah, ki so časovno pred tema dvema, se bo model učil. Pri testiranju bomo obravnavali vsako tekmo, če bo matematično upanje pozitivno, bomo hipotetično stavili na to tekmo, in če se bo naša napoved uresničila, bomo skupnemu stanju na hipotetičnem računu prišteli dobiček, v nasprotnem primeru pa bomo odšteli vložek. Izkaže se, da nam ne uspe ustvariti hipotetičnega dobička. Ob vložku enega evra na 1053 tekem v dveh sezonah naredimo 34,61 evrov izgube. Vseeno je to boljše kot povprečni uporabnik stavnic. Na eni izmed največjih svetovnih stavnic William Hill povprečni uporabnik izgubi 6,7 % svojega vložka, naš model pa je izgubil zgolj 3,3 % vložka[3].

4.3 Izboljšanje rezultatov na stavnicah

Do sedaj smo za vsako tekmo naredili enako hipotetično stavo z enakim vložkom. Model nam za vsako tekmo vrne matematično upanje in verjetnost. To lahko izkoristimo za spremembo vložka. Ko je verjetnost večja, bi bilo pametno staviti več kot takrat, ko je manjša. V drugem poizkusu vsakič stavimo zmnožek enega evra in verjetnosti za dogodek, ki ga bomo stavili, torej če bo verjetnost 90 %, bomo stavili 0,9 evra. Rezultati so precej boljši. Na 1053 tekem smo vložili 634,1 evrov in naredili zgolj 0,9 evra izgube, kar pomeni, da smo izgubili zgolj 0,1 % svojega vložka, s tem smo se precej približali cilju. Poleg verjetnosti nam model vrne še matematično upanje, ki upošteva tudi višino kvote, zato smo poizkušali še z vložki, ki so enaki matematičnemu upanju. Rezultati se tokrat ne izboljšajo, ob vložku 313 evrov izgubimo 9,3 evra, izgubili smo 2,9 % svojega vložka. S temi izboljšavami smo se precej približali svojim ciljem, a še vedno ne ustvarjamo dobička. Kot je razvidno v tabeli 3.3, verjetnost, ki jo vrne napovedni model, ni enaka dejanski verjetnosti, da se bo napovedani dogodek zgodil. Napovedni model pretirava s previsokimi verjetnostmi, porodila se nam je ideja, da

bi vrnjene verjetnosti prilagodili. Naredili smo slovar verjetnosti, ki je preslikava vrnjenih verjetnosti v klasifikacijsko točnost pri teh verjetnostih. Ključi v slovarju so vrednosti od 0 % do 95 % v intervalih po 5 %, za vsako vrednost smo zbirali število pravilno in napačno napovedanih primerov. Lahko bi se odločil tudi za krajše intervale, s tem bi dobili bolj natančen slovar, a morda bi bil preveč prilagojen specifičnim primerom. Pričakovano vrednost bomo izračunali tako, da bomo s slovarjem preslikali verjetnost. Pričakujemo lahko, da se bo tokrat model obnašal bolj konservativno. Ob vložku enega evra na 820 tekem naredimo 27,34 izgube in s tem izgubimo 3,3 %. Rezultat se ne izboljša, le manjkrat stavimo, kar je pričakovano glede na to, da smo prilagodili verjetnosti. Najbolje smo se torej odrezali, ko smo vložek prilagodili verjetnosti. Žal imamo na voljo malo podatkov, na katerih lahko testiramo hipotetične stave, zato so rezultati tega testiranja morda preveč prilagojeni podatkom. Potrebovali bi več podatkov za testiranje teh izboljšav.

Poglavje 5

Sklepne ugotovitve in nadaljnje delo

5.1 Ugotovitve in analiza

Cilj, ki smo si ga zadali na začetku diplomske naloge, ni bil dosežen, ni nam uspelo ustvariti napovednega modela, ki bi omogočal dobiček na športnih stavnicah. Vseeno nam je uspelo ustvariti model, ki je podobno točen kot najboljši napovedni modeli te vrste. Ena izmed pomembnih ugotovitev o košarki je, kako zelo pomembno je, katera ekipa ima več dni počitka. Postavlja se vprašanje, zakaj cilj ni bil dosežen. Cilj je bil zahteven, saj smo želeli doseči dobiček na stavnicah, ki imajo zaposlenih veliko ljudi, ki skrbijo, da so kvote takšne, da bodo stavnici dolgoročno prinašale dobiček. S tega vidika se zdi povsem razumljivo, da ne moremo biti uspešni proti stavnici, ki vlaga veliko denarja. Naš klasifikator je po točnosti primerljiv z najboljšimi. Vprašanje je, kako bi se najboljši klasifikatorji znašli na stavnicah, žal nismo našli takšnih podatkov.

5.2 Nadaljnje delo

Možnosti za izboljšavo je še veliko, vsi modeli pretiravajo z visokimi verjetnostmi, kar povzroča, da prevečkrat stavimo, ko je tveganje preveliko. Več algoritmov je po točnosti blizu najboljšega, morda bi boljši rezultat lahko dosegli z združevanjem le teh. Z izmero specifičnosti in senzitivnosti smo ugotovili, da dosežemo večjo točnost pri domačih zmagah. Lahko bi analizirali tekme, na katerih so zmagale gostujoče ekipe, in poskušali ugotoviti, zakaj se je to zgodilo, morda bi našli kakšne nove zanimive atribute. Atribut bi bila lahko tipična igra določene ekipe, potem bi lahko primerjali, kako določeni tipi ekip delujejo proti drugim tipom, a to bi zahtevalo veliko analize same igre košarke. Kot smo že napisali, je zelo težko premagati stavnico. Problema bi se lahko lotili povsem drugače, sedaj napovedujemo na podlagi prejšnjih podatkov o ekipi, lahko pa bi uporabili prejšnje podatke o posameznih igralcih in iz tega sestavili neko sliko o ekipi. Obstaja še veliko drugih vrst stav, kot so: število košev na tekmi, razlika v točkah, stave po četrtinah. Morda bi model, ki bi napovedoval takšne stvari bolj uspešen. V prihodnosti bi radi preizkusili svoj model na stavnici Betfair. Gre za posebne vrste stavnico, v kateri stave sklepajo uporabniki med seboj, stavna hiša pa zgolj pobere majhno provizijo od vsake sklenjene stave. Kvote niso vnaprej določene, vsak jih lahko sam določi, samo najti mora nekoga, ki je pripravljen sprejeti stavo. Na tej stavnici bi naš model moral biti pametnejši od trga, kar bi bila lažja naloga. Naš model je dosegel manjšo izgubo od povprečnega stavca, kar nas navdaja z upanjem, da bi na Betfairu lahko dosegli dobiček.

Literatura

[1] What actually is tanking, and which NBA teams actually do it?, dostopno na: http://www.sbnation.com/2014/1/10/5266770/nba-draft-lottery-tanking-gm

[2] How the San Antonio Spurs' Success Proves the Popovich Rest Method Works, dostopno na: http://bleacherreport.com/articles/2015922-how-the-san-antonio-spurs-success-proves-the-popovich-rest-method-works

[3] William Hill Annual Report and Accounts 2012, dostopno na: http://www.williamhillplc.com/~/media/Files/W/William-Hill/ar/ar_2012.pdf

[4] Ian H. Witten, Eibe Frank, Mark A. Hall, *Data mining: practical machine learning tools and techniques*,Elsevier, 2009.

[5] Anže Kravanja, *Napovedovanje zmagovalcev košarkarskih tekem*, 2012.

[6] Sean Paul Walker, *Robot Haptics*,ProQuest, 2009, str. 103.